Multilevel Models

7. Bayesian Inference in GLMMs

Germán Rodríguez

Princeton University

April 16, 2018

Generalized linear multilevel models

The logit models we have discussed are special cases of the generalized linear mixed/multilevel model. We assume that conditional on a set of multivariate normal random effects

$$\mathbf{u} \sim N_q(\mathbf{0}, \boldsymbol{\Omega})$$

the outcome **y** has a distribution in the *exponential family*, which includes the normal, binomial, Poisson, gamma and others. We further assume that the conditional expectation satisfies

$$E(\mathbf{y}|\mathbf{u}) = f^{-1}(\mathbf{X}\beta + \mathbf{Z}\mathbf{u})$$

where **X** is the model matrix for the fixed effects β, **Z** is the model matrix for the random effects **u**, and $f()$ is a one-to-one transformation called the *link* function, which includes the identity, logit, probit, c-log-log, log, and others.

The marginal likelihood has a closed form for normal models with identity link, but otherwise involves intractable integrals.

Maximum likelihood estimation

ML estimates can be computed by numerical integration of the likelihood function using Gaussian quadrature, but the procedure is computationally intensive and can only be used for simple models.

A two-level random-intercept logit or poisson model requires a one-dimensional integral. Using 12 quadrature points is equivalent to 12 logit or Poisson likelihoods.

A three-level random intercept model, or a two-level model with a random intercept and slope, requires a two-dimensional integral. One evaluation of a logit or Poisson likelihood using 12 quadrature points per level is equivalent to 144 one-level models.

A three-level logit model with two random coefficients per level using 12-point quadrature for each, is equivalent to evaluating almost 21,000 logit or Poisson likelihoods. Numerical integration doesn't scale well, and soon succumbs to the "curse of dimensionality".

Bayesian estimation

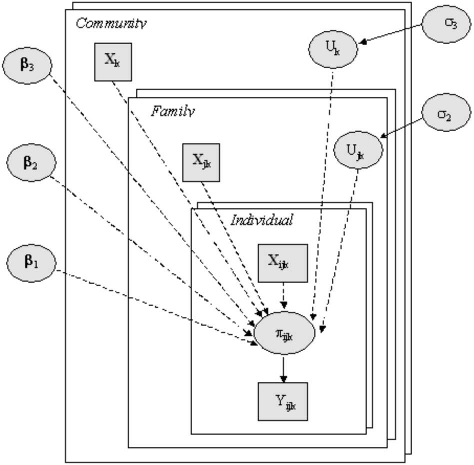
Recent advances in Bayesian estimation avoid the need for numerical integration by taking repeated samples from the posterior distribution of the parameters of interest.

To apply this framework we adopt a Bayesian perspective, treating all parameters as random variables and assigning prior (or hyperprior) distributions to the fixed parameters $\boldsymbol{\beta}$ and to the variances $\mathbf{\Omega}$ of the random effects. (We have, of course, already assigned a prior distribution to the random effects $\mathbf{u}$.)

To obtain Bayesian estimates that are roughly comparable to maximum likelihood estimates, many analysts use vague or non-informative priors. Fixed effects are typically assumed to come from normal distributions with mean zero and very large variances. Precisions, defined as the reciprocals of variances, are often sampled from diffuse gamma distributions. Gelman suggests using a uniform prior on the standard deviation instead.

A graphical model

Bayesian models are often shown in graphical form as illustrated below for a 3-level random intercept logit model



Here we need priors for the βs and hyperpriors for the σs.

The Gibbs sampler

A popular method for drawing observations from a posterior is the Gibbs sampler, which draws from a joint distribution by sampling repeatedly from each of the full conditional distributions.
In our case we need to sample from the joint distribution of the parameters given the data, which we write as

$$[\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{u} | \mathbf{y}]$$

The Gibbs sampler tells us that we can sample instead from the three full-conditionals

$$[\boldsymbol{\beta} | \boldsymbol{\sigma}^2, \mathbf{u}, \mathbf{y}], \quad [\boldsymbol{\sigma}^2 | \boldsymbol{\beta}, \mathbf{u}, \mathbf{y}], \quad \text{and} \quad [\mathbf{u} | \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{y}]$$

which in our case further simplify to

$$[\boldsymbol{\beta} | \mathbf{u}, \mathbf{y}], \quad [\boldsymbol{\sigma}^2 | \mathbf{u}] \quad \text{and} \quad [\mathbf{u} | \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{y}]$$

The fixed effects depend only on the random effects and response, and the variances depend only on the random effects.

Markov chains

Let $\boldsymbol{\beta}_k, \boldsymbol{\sigma}^2_k$, and $\mathbf{u}_k$ denote a sample, with $k = 0$ representing initial values of the fixed and random parameters. The Gibbs sampler draws

$$\begin{array}{lll} \boldsymbol{\beta}_{k+1} & \text{from} & [\boldsymbol{\beta}|\mathbf{u}_k, \mathbf{y}] \\ \boldsymbol{\sigma}^2_{k+1} & \text{from} & [\boldsymbol{\sigma}^2|\mathbf{u}_k] \quad \text{and} \\ \mathbf{u}_{k+1} & \text{from} & [\mathbf{u}|\boldsymbol{\beta}_{k+1}, \boldsymbol{\sigma}^2_{k+1}, \mathbf{y}] \end{array}$$

This is a Markov chain because each sample depends only on the previous one. Under reasonably general conditions, the sample converges in distribution to the joint posterior of interest.

Usually one discards a “burn-in” period long enough to ensure that the chain has converged to its stationary distribution and uses the remaining observations to estimate features of the posterior, such as the mean or a credible interval.

The sample is **not** i.i.d. The efficiency of the chain is lower when the draws are highly correlated.

Sampling methods

The actual sampling is done using methods appropriate for each distribution. I'll mention just a couple of approaches.

Uniform Distribution. An indispensable starting point is a routine to generate pseudo-random numbers or samples from the uniform distribution in $(0,1)$, which both Stata and R do well. `runiform` in Stata.

The Inversion Method. A useful general method is based on the fact that

$$\text{if } X \sim F(x) \quad \text{then} \quad F(X) \sim U(0,1)$$

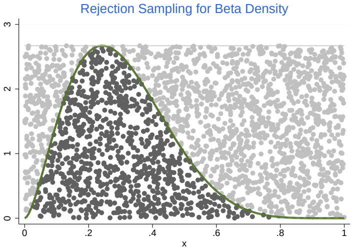
If we can invert the c.d.f. we can then draw samples from it by calculating $F^{-1}(u)$ where $u \sim U(0,1)$.

For example we could draw normals this way. But Stata and R have specialized function for many distributions including beta, binomial, χ^2, gamma, hypergeometric, normal, Poisson, and

Rejection sampling

What if the distribution you need, say $f(x)$, isn't in the list? There's an ingenious method called rejection (or importance) sampling which has wide applicability.

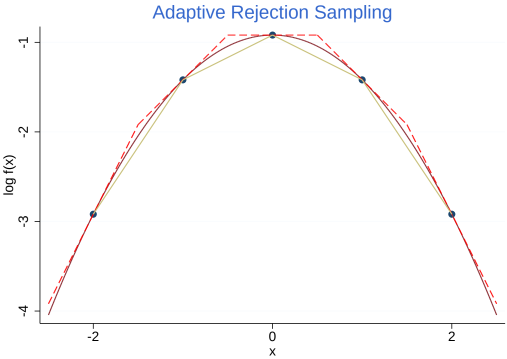
All you need is another density that you know how to sample from, say $g(x)$, which "covers" $f(x)$ in the sense that it has the same domain and there is a constant c such that $cg(x) \geq f(x)$ for all x.



You then draw a sample x from $g(x)$ and keep it with probability $f/cg(x)$, which you do by sampling u from $U(0,1)$ and comparing it to the ratio above. This corrects for the fact that sampling from $g(x)$ oversamples values of x where $g(x)$ is "taller" than $f(x)$.

Adaptive rejection sampling (ARS)

Gilks proposed a sampling method that extends rejection sampling to any log-concave distribution, using outer and inner envelopes based on $f(x)$ and its derivative $f'(x)$ at selected points.



The method samples from the outer envelope, accepts values under the inner envelope, and otherwise evaluates $f(x)$ to decide and then $f'(x)$ to tighten the envelopes.

WinBUGS

Gibbs sampling using adaptive rejection sampling has been implemented in a package called BUGS (Bayesian Inference Using the Gibbs Sampler). The windows version is called WinBUGS.

Two alternatives are OpenBUGS and JAGS (Just Another Gibbs Sampler), as well as MLwIN, but we'll focus on WinBUGS. The program lets you describe your model using a declarative language to specify the prior distributions and the likelihood, and uses an expert system to derive the posterior and decide whether to use a specialized sampling method or ARS. We'll consider two examples:

1. An analysis of immunization in Guatemala based on a three-level random-intercept model as illustrated in slide 5, comparing results with other methods.
2. A study of hospital delivery data from Lillard and Panis.

In the process we will address the important issue of convergence diagnostics.

Immunization in Guatemala

Here are results from the immunization model described in RG2

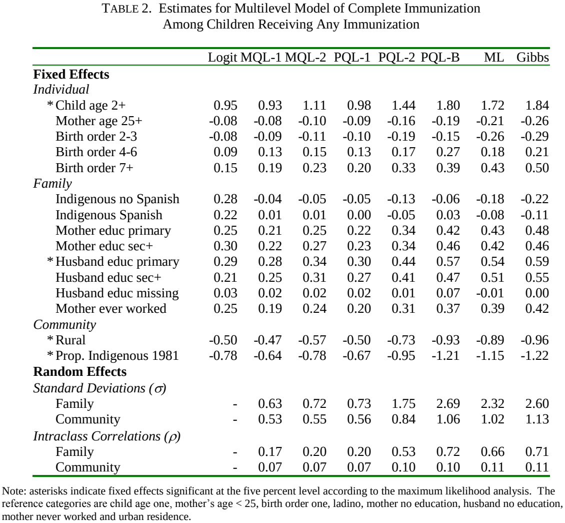


TABLE 2. Estimates for Multilevel Model of Complete Immunization Among Children Receiving Any Immunization

	Logit	MQL-1	MQL-2	PQL-1	PQL-2	PQL-B	ML	Gibbs
Fixed Effects								
Individual								
*Child age 2+	0.95	0.93	1.11	0.98	1.44	1.80	1.72	1.84
Mother age 25+	-0.08	-0.08	-0.10	-0.09	-0.16	-0.19	-0.21	-0.26
Birth order 2-3	-0.08	-0.09	-0.11	-0.10	-0.19	-0.15	-0.26	-0.29
Birth order 4-6	0.09	0.13	0.15	0.13	0.17	0.27	0.18	0.21
Birth order 7+	0.15	0.19	0.23	0.20	0.33	0.39	0.43	0.50
Family								
Indigenous no Spanish	0.28	-0.04	-0.05	-0.05	-0.13	-0.06	-0.18	-0.22
Indigenous Spanish	0.22	0.01	0.01	0.00	-0.05	0.03	-0.08	-0.11
Mother educ primary	0.25	0.21	0.25	0.22	0.34	0.42	0.43	0.48
Mother educ sec+	0.30	0.22	0.27	0.23	0.34	0.46	0.42	0.46
*Husband educ primary	0.29	0.28	0.34	0.30	0.44	0.57	0.54	0.59
Husband educ sec+	0.21	0.25	0.31	0.27	0.41	0.47	0.51	0.55
Husband educ missing	0.03	0.02	0.02	0.02	0.01	0.07	-0.01	0.00
Mother ever worked	0.25	0.19	0.24	0.20	0.31	0.37	0.39	0.42
Community								
*Rural	-0.50	-0.47	-0.57	-0.50	-0.73	-0.93	-0.89	-0.96
*Prop. Indigenous 1981	-0.78	-0.64	-0.78	-0.67	-0.95	-1.21	-1.15	-1.22
Random Effects								
Standard Deviations (σ)								
Family	-	0.63	0.72	0.73	1.75	2.69	2.32	2.60
Community	-	0.53	0.55	0.56	0.84	1.06	1.02	1.13
Intraclass Correlations (ρ)								
Family	-	0.17	0.20	0.20	0.53	0.72	0.66	0.71
Community	-	0.07	0.07	0.07	0.10	0.10	0.11	0.11

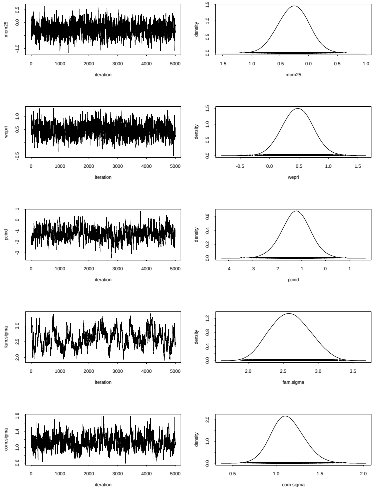
Note: asterisks indicate fixed effects significant at the five percent level according to the maximum likelihood analysis. The reference categories are child age one, mother's age < 25, birth order one, ladino, mother no education, husband no education, mother never worked and urban residence.

Trace plots and posteriors

Here is what the actual output looks like for selected parameters.

The first three parameters are child, mother and community fixed effects, the last two are standard deviations of mother and community random effects.

On the left we see trace plots, which ideally should look like fuzzy caterpillars, but the family σ shows slow mixing. On the right we see kernel estimates of the posterior densities.



Convergence diagnostics

Some Bayesians recommend using several chains and others prefer one long chain. A good compromise is to run three chains with different starting points.

For the Guatemala data we ran burn-ins of 200 followed by 5,000 draws. A battery of tests showed that this was adequate, based on

Geweke (1992)'s test of convergence, which divides the chain into two sections (such as first 10% and last 50%) and compares means

Raftery and Lewis (1992, 1996) `gibbsit` software, to determine the sample size needed to estimate each posterior c.d.f. at 95% credible limits within 0.015 with probability 0.95

Roberts (1996) estimate of efficiency, to ensure that we had the equivalent of at least 100 i.i.d. observations in the worst case, where efficiency was only 2%.

Convergence diagnostics

There's a specialized package for convergence diagnostics and output analysis for Gibbs output called CODA. The ecology is richer in the R world than in Stata, but see Thompson, Palmer and Moreno (2006) in the Stata Journal 6:530-549, available at http://www.stata-journal.com/sjpdf.html?articlenum=st0115

The website illustrates the use of winBUGS to estimate a logit model for the hospital delivery data. This model is simple enough that it can be fitted by maximum likelihood using Stata or R, but it is instructive to try the Bayesian approach, which gives similar results if we use non-informative priors.

The pages at hospBUGS.html and hospBUGS2.html have step-by-step instructions for running the model using the GUI with a compound document, and using the scripting facility introduced with version 1.4. To get winBUGS visit http://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/.